Title

Replication and Extension of the Comparison of Wrist Actigraphy and Sleep Diaries in Measuring Sleep in Healthy Adults

Authors

Asatryan, Sona
Azarmi-Por, Ryan
Contreras, Eli
et al.

Publication Date

2024-04-23

Replication and Extension of the Comparison of Wrist Actigraphy and Sleep Diaries in Measuring Sleep in Healthy Adults

Grace Judge, Sona Asatryan, Ryan Azarmi-Por, Eli Contreras, Henry Lin, Devan Melwani, Sharon Vuong

Cognitive Science and Psychology Undergraduate Laboratory at Berkeley

Undergraduate Student Mentor: Grace Judge

Graduate Student Mentor: Colin Jacobs

University of California, Berkeley

Abstract

Quality sleep is highly indicative of human functionality and health, with links to restoration and healing. The ability to accurately measure sleep quality is especially important for disease diagnosis, treatment development, and sleep pattern monitoring. Clinical studies and healthcare facilities utilize quantitative and qualitative methods, namely wrist actigraphy and sleep diaries, to measure sleep quality. This paper replicates and extends data collected from a previous study comparing wrist actigraphy and sleep diary measurements of sleep in healthy adults (Alyobi & Sherman, 2021). In our replication of Alyobi and Sherman's statistical analysis, including mean and standard deviation calculation, t-tests, and Pearson's Correlation Coefficient (PCC) tests, our overall results were consistent with the findings of Alyobi and Sherman, that wrist actigraphy and sleep diary recordings were similar and differed only on a day by day basis. To extend the given data, we conducted an F test to measure variance and calculated the median and Interquartile Range (IQR) for each measurement across the five days. Additionally, we constructed a violin plot to alternatively compare the wrist actigraphy and sleep diary measurements. We found that both methods of recording sleep still only varied on a day-by-day basis, with no significant variance between the two methods. We also concluded that the data set was not vulnerable to outliers or a heterogeneous spread of values, with most values grouped closely together, giving greater strength to Alyobi and Sherman's findings and subsequent conclusions.

REPLICATION AND EXTENSION OF THE COMPARISON OF WRIST ACTIGRAPHY AND SLEEP DIARIES IN MEASURING SLEEP IN HEALTHY ADULTS

Introduction

Sleep is a core component of human functionality and health, and accurate measurements of sleep duration and quality are thus vital for many areas of research (Adam et al., 1984). Polysomnography (PSG) is considered the standard for detecting interruptions during sleep, but it has limitations. One major limitation is that PSG is often unsuitable for long-term or longitudinal studies because the method has high cost and participant burden (Lehrer et al., 2022). Sleep actigraphy and self-reported sleep diaries are alternatives to PSG that are considered cost-effective and easy to implement.

Wrist actigraphy measurements employ an actigraph machine, which uses a piezoelectric accelerometer to detect movement. The subject simply needs to wear this device on their wrist, similar to a watch, and the device will record the subject's movement continuously for an extended period of time. Actigraphy can measure a subject's Total Sleep Time (TST), Sleep Latency (SL), the time it takes for them to fall asleep, and Wake After Sleep Onset (WASO), the extent to which sleep was disrupted after the participant had fallen asleep.

In addition, subjective sleep diaries are another method of measuring sleep quality. The participant writes about their previous night's sleep, recording whether they woke up, the number of times they woke up, how long they slept, etc. Sleep diaries are widely considered a more accessible and significantly less expensive alternative to PSG and actigraphy. Still, actigraphy and sleep diaries have been noted to provide individual differences in data (Thurman et al., 2018).

So which method is more effective? To start, wrist actigraphy seems to be an effective, psyiological method. However, studies have shown that wrist actigraphy measurements might not be as reliable as many claim it to be. Numerous studies have revealed that the device incorrectly measures TST and WASO. Maich et al. (2018) found that wrist actigraphy devices overestimated TST measurements compared to self-reported sleep diaries. However, studies done by Kawada et al. (2023) found that actigraphy underestimated TST compared to PSG, raising concerns about actigraphy's effectiveness. On the other hand, there is uncertainty regarding the accuracy of sleep diary measurements as well. Researchers have found that compliance with sleep diaries becomes less consistent over longer periods of time. Thurman et al. (2018) compared the effectiveness of both wrist actigraphy and sleep diaries in measuring sleep quality and quantity among 30 healthy adults over a period of 16 weeks. They concluded that wrist actigraphy reported a shorter TST than sleep diary. In addition, Thurman et al. (2018) concluded that 87% of subjects completed their sleep diaries at the beginning of the experiment, whereas the percentage had decreased to 67% at the end of the experiment. Their conclusions highlight the need for further investigation into the efficacy and variation of wrist actigraphy and sleep diary measurements.

Overall, both wrist actigraphy and sleep diaries are promising tools for measuring sleep quality. However, the aforementioned discrepancies between wrist actigraphy and sleep diary measurements necessitate further investigation into the accuracy and efficiency of both methods. In their 2021 paper “A direct comparison of wrist actigraphy and sleep diaries in measuring sleep in healthy adults”, researchers Alyobi and Sherman set out to explore the differences between wrist actigraphy and sleep diaries in measuring sleep and sleep quality. In their investigation, Alyobi and Sherman measured TST in young healthy adults over a period of 5 consecutive days, and concluded that there was no significant difference in wrist actigraphy and sleep diary measurements, finding only day-by-day variation in sleep. This paper will replicate the aforementioned results, and statistically expand upon them for additional information regarding wrist actigraphy and sleep diary measurements of sleep.

Methods

Participants were recruited for Alyoobi and Sherman's larger study on sleep memory; a preliminary experiment was conducted to compare wrist actigraphy and sleep diary measurements. Sixty-eight participants between ages 18 and 65 were recruited, consisting of 52 females and 16 males, who were either given one hour of course credit or compensated 10 pounds(£). TST sleep data was collected by the Motion Watch8 actigraph machine, which was worn on the non-dominant hand. Participants pressed the market button when they went to sleep and woke up. Participants also recorded sleep duration in a sleep diary by recording the times they went to sleep and awoke. Participants were given detailed instructions on how to both wear the watch and record data in their sleep diary. Data was recorded for 5 consecutive days.

Replication Data Analysis

Mean Total Sleep Time and Standard Deviation

For each day of the experiment, we replicated Alyobi and Sherman's calculation of the mean Total Sleep Time (TST) and Standard Deviation (SD) of the wrist actigraphy and the sleep diary measurements independently. These calculations were conducted in Excel, and shown in Table 1 in the results section.

T-test

We replicated the t-test assessing whether the differences between the wrist actigraphy and sleep diary data were statistically significant using R programming. The t-statistic value itself measures the difference between the means of the wrist actigraphy and the sleep diary method relative to the variability in both groups. The larger the t-statistic, the larger the difference between the means. The t-test also provides a p-value that measures the probability of observing any difference between the means if the null hypothesis (there is no statistical difference between groups) were considered true.

By separating the diary and actigraphy data for each specific day and utilizing the t.test function in R, we successfully repeated t-tests for all five days and compared them with the original findings, as shown in Table 2 of the results section.

Pearson's Correlation Coefficient

We replicated Pearson’s Correlation Coefficient using the cor.test function in R in comparison to the values presented by Alyobi and Sherman, finding the correlation coefficient for each day of wrist actigraphy and sleep diary measurements. The Pearson’s Correlation Coefficient (PCC) test gives us a correlation coefficient, r, between -1 and 1. 1 would be a strong positive correlation, -1 would be a strong negative correlation, and 0 would be no correlation. The r values are reported in Table 3 below and compared with the original results.

Extension Data Analysis

F-test

For our extension of the findings of Alyobi and Sherman (2021), we employed the F-test using the var.test function in R to measure the significance of the degree of variation between the wrist actigraphy measurements and the sleep diary measurements. The null hypothesis of the F-test assumes that the two groups have no significant variation. The F-value represents the ratio of variances between the two populations, the larger variance being a number further from 1. The F-test also provides a p-value, indicating the probability of obtaining that F-value under the null hypothesis that any variance between the two populations is insignificant and due to chance. Upon replicating the t-test to measure the difference of the means, we decided to expand the data with a F-test to measure the variance between the groups and determine whether it is significant. While the t-test only provides information about means, the F-test gave us additional insight into the homogeneity and variance of the data groups.

Interquartile Range (IQR) and Median

In our extension, we also decided to utilize the Median and Interquartile Range (IQR) because they provide us with measurements of central tendency and dispersions, particularly investigating skewness or outliers in the data. First, we calculated the median, which represents the middle value of the dataset when arranged in ascending order. Using the median on certain occasions can fact-check if the mean values are correct, as the mean cannot account for outliers or skewness in the data. Overall, the median provided us with a measure of central tendency less susceptible to outliers compared to the mean. Furthermore, to understand the spread of the data, we utilized the interquartile range (IQR), which quantifies the dispersion of the central half of the dataset. To calculate the IQR, we determined the values corresponding to the first quartile (Q1) and the third quartile (Q3), representing the 25th and 75th percentiles, respectively. Then, we subtracted Q1 from Q3 to obtain the IQR. This measure is particularly useful as it is less affected by extreme or outlier values compared to the range. The median and IQR for the watch and dairy measurements across the 5 days were calculated.

Violin Plot

We chose to construct a violin plot to extend the comparisons between wrist actigraphy and sleep diary measurements, using R programming with the plotData, ggplot(), and geom_violin() functions. Violin plots are advantageous in that they allow us to compare the overall shape and density of the distribution of data, and provide a sense of both range and variance of the data through qualitative observation.

Results

Replication

Table 1. Comparison of means and standard deviations of daily actigraphy and sleep diary data

	M ± SD (N=68)			
	Actigraphy (replicated)	Actigraphy (presented by Alyobi and Sherman, 2021)	Sleep Diary (replicated)	Sleep Diary (presented by Alyobi and Sherman, 2021)
DAY 1	488 ± 117	488 ± 117	498 ± 119	497 ± 118

DAY 2	475 ± 102	475 ± 102	496 ± 91	496 ± 91
DAY 3	513 ± 95	515 ± 96	552 ± 112	552 ± 112
DAY 4	496 ± 98	493 ± 100	512 ± 112	512 ± 112
DAY 5	456 ± 89	456 ± 89	483 ± 109	485 ± 110
OVERALL	489 ± 54	490 ± 54	508 ± 58	507 ± 58

As seen in Table 1, we found that overall mean values differed by a minute less than presented for diary data and a minute more than presented for actigraphy data(i.e. Day 1 Diary M = 498, SD = 119, Day 1 Watch M = 497, SD = 118). Still, there was no overall change in standard deviation. On Day 1, there was a slight variation in sleep diary values; replicated means results for sleep diary presented one value higher than presented results, affecting standard deviation calculations. Day 2 was constant with replicated and presented values, while Day 3 actigraphy values also appeared to have slight variation. On Day 4, the replicated value for actigraphy was 2 minutes higher than the value presented, causing the standard deviation to decrease as well. Lastly on day 5, replicated values of the sleep diary were 2 minutes lower than the presented value. The slight differences in our calculated values and the presented values by Alyobi and Sherman are possibly due to the unreported exclusion of specific outliers or cases in their calculations.

Table 2. Comparison of t-test t-values and p-values

Day	Replicative t-value	Replicative t-value	t-value % Difference (Difference/ Average)	Replicative p-value	Replicative p-value	p-value %Difference (Difference/ Average)
1	-0.857	-0.85	0%	0.394	0.394	0%
2	-1.845	-1.84	0%	0.069	0.069	0%
3	-3.064	-2.91	5.25%	0.003 (<0.01)	<0.01	0%
4	-1.380	-1.38	0%	0.172	0.172	0%
5	-2.759	-2.95	6.69%	0.007 (<0.01)	<0.01	0%

Our replicated values from the t-test show that there is no significant difference between the wrist and diary measurements of sleep during days 1, 2, and 4 of the experiment. Specifically, comparisons showed t(67), -.857, p=.394 for day 1, t(67), -1.845, p=.069 for day 2, and t(67), -1.38, p=.172 for day 4 of the experiment. On the other hand, our replicated results showed significant differences for days 3 t(67), -3.064, p=.003 and 5 t(67), -2.759, p=.007. Since the replicated and presented p-values for Day 3 (p=0.003) and 5 (p=0.007) were both less than 0.01, we can reject the null hypothesis and conclude that there is a statistically significant difference between the two methods for these two days, with the diary recording a significantly greater mean TST than actigraphy (485 ± 110 and 456± 89 respectively), consistent with the findings of Alyobi and Sherman.

Table 3. Comparison of PCC values

Day	Replicated PCC	Presented PCC	% Difference (difference/average)
1	0.68	0.68	0%
2	0.53	0.53	0%
3	0.50	0.50	0%
4	0.57	0.57	0%
5	0.68	0.70	2.9%

By comparison, the replicated Pearson's Correlation Coefficient (PCC) has similar values presented by Alyobi and Sherman. Days 1 through 4 represent no difference between replicated and presented values with zero percent difference. However, day 5 has a 2.9 percent difference between the replicated PCC value (0.68) and the represented value (0.70). The discrepancy between these values may be due to the use of a full data set while Alyobi and Sherman may have removed any outliers in the process, similar to the slight differences we found in our mean, standard deviation, and t-test calculations.

Extension

The F-tests performed for all days of the experiment indicate there is no significant variance between the groups (Table 4). With a significance level of .05, we determined that the lower critical value is 1.65, and the upper critical value is .63. As seen in Figure 5, the F-values for all days of the experiment fall between the range of critical values, so we concur with the null hypothesis that there is no significant difference in the variance between the two groups. Additionally, all p-values exceeded the .05 threshold, reassuring concurrence with the null hypothesis that there is no significant difference between the degree of variance of the groups.

Table 4. F-test extension data

Day	F-Value	p-value
1	0.98069	0.9366
2	1.2634	0.3409
3	0.74093	0.2223
4	0.79162	0.3412
5	0.68606	0.1256

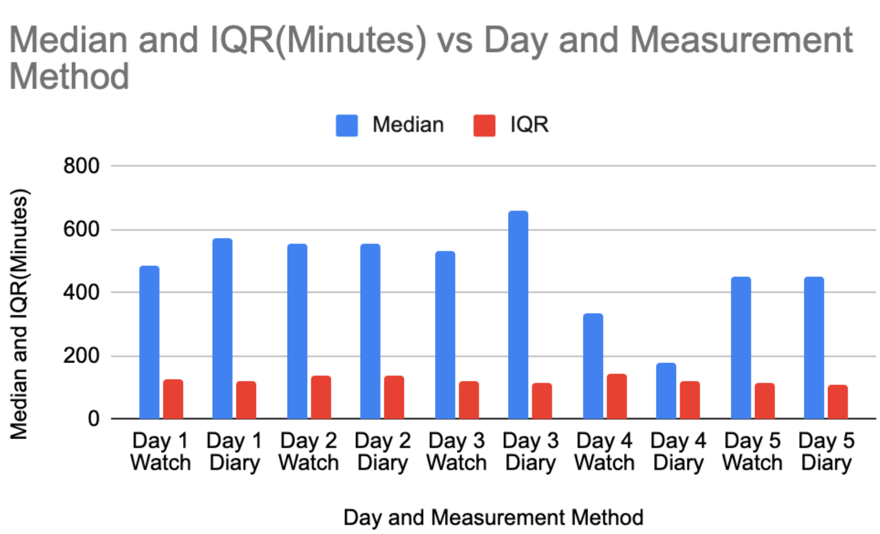
In our extension of Alyobi and Sherman's data, our calculated median values were very similar to the mean values, ranging from 392 to 575. The IQR results showed Day 4 had the greatest variability in values with a watch IQR of 140.25 and 122.5. IQR calculations report no skewness between wrist actigraphy and sleep diary measurements. These results are reported in Table 5, and visualized in Figure 1.

Table 5. Median/IQR

	Median	Q1	Q3	IQR

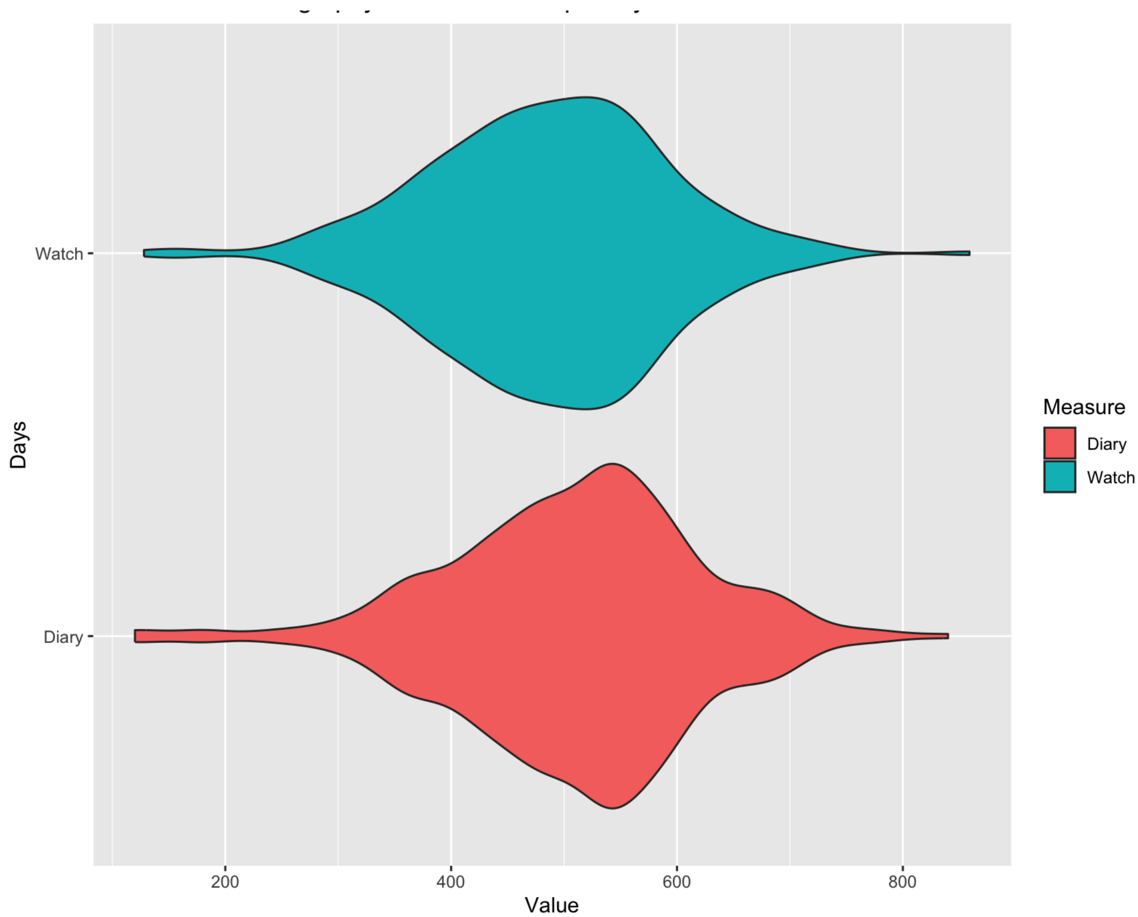
Day 1 Watch	484.00	425.00	548.50	123.50
Day 1 Diary	472.50	447.50	570.00	122.50
Day 2 Watch	563.50	402.50	538.50	136.00
Day 2 Diary	540.00	420.00	558.75	138.75
Day 3 Watch	489.00	444.75	563.25	118.50
Day 3 Diary	575.00	483.75	600.00	116.25
Day 4 Watch	501.50	428.75	569.00	140.25
Day 4 Diary	450.00	450.00	572.50	122.50
Day 5 Watch	392.50	396.50	511.25	114.75
Day 5 Diary	415.00	429.00	540.00	111.00

Figure 1: Median and IQR Bar Chart



As part of our extension, we created a violin plot to visualize a combination of all of the TST recordings from all five days. We constructed one plot for wrist actigraphy and one for sleep diary results, as seen below in Figure 2.

Figure 2: Violin Plot of Wrist Actigraphy and Sleep Diary Data



Discussion

Upon replication of Alyobi and Sherman's 2021 paper "A direct comparison of wrist actigraphy and sleep diaries in measuring sleep in healthy adults", our replicated values of mean TST and SD, T-test, and PCC were observed to be consistent with the reported values, with slight discrepancies. Due to the small magnitude of variation, we attribute these differences to excluded or modified data in the original study, as we only worked with the raw data provided.

As stated in the results section, wrist actigraphy measurements generally reported a lower overall TST relative to subjective sleep diary measurements. Both the replicated and presented differences between the wrist actigraphy and sleep diary scores are not consistently different: the t-test replication showed no significant difference in TST scores between the actigraphy and diary on the first, second, and fourth days, but determined a significant difference on the third and fifth days. In other words, there is no significance in the difference in measured sleep time of the wrist actigraph and sleep diary on the first, second, and fourth day. However, on the third and fifth days, the measured sleep time for the two methods showed different times, which likely suggests some sort of discrepancy among the participants when using the two methods to measure sleep. The difference in the t-test replication on the third and fifth days shows potential inconsistencies in participants' reporting when utilizing the two methods to assess sleep. This discrepancy implies that certain factors might have influenced participants' ability to accurately record their sleep duration, particularly on these specific days. Therefore, while the lack of significant differences on the first, second, and fourth days indicates a general agreement between the two measurement methods, the observed disparities on the third and fifth days warrant further investigation into the factors influencing participants' reporting accuracy.

Overall, our replicated t-test values of wrist actigraphy and sleep diary TST means align with the results of Alyobi and Sherman, indicating variation in measurements is caused by day-by-day variation in sleep and is likely not dependent on the type of measurement device. Our Pearson's Correlation Coefficient calculations also concurred with Alyobi and Sherman's conclusions of a moderate to strong correlation between measurements of wrist actigraphy and sleep diary data, the strength of which also varied day to day.

Our extension of Alyobi and Sherman's 2021 paper was rooted in understanding of the variance and range in the dataset, as well as adjusting the analysis for potential outliers. The standard deviation was already calculated, and considering it as the square root of variance, we could already see that there were indeed different values of variance on a day-by-day basis. We used the F test to contextualize this with the methods of sleep recording and found that there is no significant difference in variance between measurements of actigraphy and sleep diary across the five days of the study.

To address some of the potential discrepancies in the data due to outliers or having a wide range of values, we calculated the median and the IQR for the sleep diary and wrist actigraphy data for each day. As mentioned previously, the median and mean results were very similar, and the IQR concluded that the measurements are relatively homogenous and suggest no skewness. We can therefore determine that the data analysis was not particularly affected by outliers or a wide range of values. In other words, our replication utilization of the median and the IQR shows us that there are no significant outliers between the measured time of the wrist actigraphy and the sleep diaries. Our constructed violin plot stayed consistent with our overall findings, as we can visually see that the data is closely grouped in the characteristic violin shape, without many bulges or irregularities representing major outliers or a broad spread of data. We also noted that the plot for the sleep diary data seemed more “bumpy” than the plot for the wrist actigraphy data, which could suggest some heterogeneity in the subjective measurements of the diary. However, in consideration of our F test results, we determined that there is indeed no significant variance between the two methods of measurement. We conclude that the shape of the violin plot data between wrist actigraphy and sleep diary recording stays largely consistent, and any differences can be attributed to day-by-day variation in sleep.

The replication portion of our paper implies a low degree of variability between the wrist actigraphy and sleep diary as modes of measurement of sleep and sleep quality. Upon further expansion of Alyobi and Sherman’s data, our aforementioned results strongly imply there is no significant variation between wrist actigraphy and sleep diary measurements, despite the differences observed in days 3 and 5. The accuracy of a sleep diary is largely contingent on the individual’s ability to assess their own sleep quality. When only two out of the five days show significant differences, it is plausible to assume that each participant’s unique assessment of sleep quality evokes inconsistencies between both methods of sleep analysis. Furthermore, the study was only conducted across five days. This is a considerably short timeframe to confidently glean any concrete determination about the efficacy of either device. To conclude whether the variation was due to the true capabilities of each recording device or simply just study design, it would be useful to replicate this study over the span of at least a month. Essentially, we can suggest that both sleep diaries and wrist actigraphy are comparably accurate tools to measure sleep and sleep quality. These two methods of sleep analysis are comparable, and any variation in their recordings in the long term could be attributed to the individual's sleep or ability to record their sleep, rather than the machine's measurement.

However, our calculations and statistical analysis are based solely upon the paper's raw data, thus posing a limitation to our extension of their data. As aforementioned, minor discrepancies in our data replication could be attributed to changes made in the final data set during statistical analysis. Having this information would be useful in explaining the differences between the presented data and our replicated data. Furthermore, Alyobi and Sherman's raw data lacks demographic information, like sex, age, family status, and more. Demographics could severely affect TST measurements and introduce an interesting avenue for future exploration regarding demographics and TST. Finally, another limitation of the research of Alyobi and Sherman(2021) is the lack of reference to a third "gold standard" level of reporting (i.e. PSG). They noted that although there are differences between the actigraphy and diary reporting scores generated, without reference to a "gold standard" measurement, it is not possible to say which of these tools was the most accurate. A possible extension of this study could bring in a "gold standard" level of reporting to determine whether the sleep diary or wrist actigraphy mode of measurement is the most accurate and efficient tool to measure sleep.

Conclusion

Overall, our replication and extension of the study compared the total sleep times measured by wrist actigraphy and sleep diaries as methods across five consecutive days in healthy young adults. Our replication results support the conclusions of Alyobi and Sherman (2021), with minor discrepancies in the statistical analyses. Overall, the replication analysis indicated significant differences only between the measure of total sleep times with the wrist actigraphy and sleep diary measurements on the third and fifth day of the experiment. This result is consistent with Alyobi and Sherman's findings, and can be attributed to day-to-day variability. For our extension, we utilized various methods including F-tests, median, interquartile range (IQR), and violin plots, verifying that there is no significant variability between the two measurement methods and that there were no significant outliers within the dataset. Our extension thus further supported the results of Alyobi and Sherman, suggesting comparable sleep measurements between the two methods across the five days. Overall, our extension highlighted the effectiveness of both wrist actigraphy and sleep diaries as methods of measuring sleep and assessing sleep quality in healthy individuals. Our conclusions have important implications in medical settings and clinical research using sleep measurements as an important parameter. Both wrist actigraphy and sleep diaries will result in similar and accurate measurements of total sleep time. This could be used in a wide range of settings and could have many implications. For instance, these tools could be used by medical professionals monitoring their patients' recovery, or clinical research participants in sleep-related disorder studies. Research on total sleep time can contribute to our understanding of sleep disorders by identifying diagnostic criteria, prevalence rates, and potential risk factors. We concluded that the wrist actigraphy and sleep diary measurements as equally accurate and important tools for measuring sleep, with important implications in the health setting. However, further studies are needed to validate these findings and explore additional methods for optimizing sleep measurement techniques.

Acknowledgements

Our ULAB group would like to thank Professor D'esposito for his continual support and encouragement of the ULAB program. We would also like to thank Colin Jacobs as our graduate student advisor for our replication and extension of this study.

References

1. Adam, K, and I Oswald. "Sleep helps healing." BMJ, vol. 289, no. 6456, 24 Nov. 1984, pp. 1400–1401, https://doi.org/10.1136/bmj.289.6456.1400.
2. Alyobi, Zainab, and Susan M Sherman. "A Direct Comparison of Wrist Actigraphy and Sleep Diaries in Measuring Sleep in Healthy Adults.", 19 Nov. 2021, https://doi.org/10.31234/osf.io/3uzwg.
3. Kawada, Tomoyuki. "Poor agreement between actigraphy and polysomnography for assessing sleep during post-traumatic amnesia." Journal of Clinical Sleep Medicine, vol. 19, no. 1, Jan. 2023, pp. 201–201, https://doi.org/10.5664/jcsm.10318.
4. Lehrer, H Matthew, et al. "Comparing polysomnography, actigraphy, and sleep diary in the home environment: The study of women's health across the nation (swan) sleep study." SLEEP Advances, vol. 3, no. 1, 1 Jan. 2022, https://doi.org/10.1093/sleepadvances/zpac001.
5. Maich, Kristin H., et al. "Psychometric Properties of the consensus sleep diary in those with insomnia disorder." Behavioral Sleep Medicine, vol. 16, no. 2, 27 May 2016, pp. 117–134, https://doi.org/10.1080/15402002.2016.1173556.
6. Thurman, Steven M., et al. "Individual differences in compliance and agreement for sleep logs and wrist actigraphy: A longitudinal study of naturalistic sleep in healthy adults." PLOS ONE, vol. 13, no. 1, 29 Jan. 2018, https://doi.org/10.1371/journal.pone.0191883.